Forecasting Stock Prices with Machine Learning: A Practice in China A-Share Market

Yikun Jiang
Jinan University, Guangzhou, China

Abstract: A supervised machine learning method, Support Vector Machine (SVM) has been employed in this study to predict the stock prices of Kweichow Moutai (Maotai) and Contemporary Amperex Technology (CATL) in China by using daily trading data, macroeconomic indicators and events, seasonal impact as well. For each stock, three models were developed based on different features included: (1) stock-specific features only, (2) adding macroeconomic factors, and (3) including seasonal indicators and global events. Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Direction Accuracy were used to assess model performance. Results indicate that including macroeconomic factors improved predictive accuracy for CATL only.

Keywords: Machine Learning; Stock Market; SVM; Regression.

1. Introduction

For the top 5 stock exchange markets, the daily trading volume of each is over 10 billion USD, according to the daily summary up to end of June 2024. As the game center of money and fortune, stock market prediction has always been a fundamental topic of interest in finance and investment for decades. Since financial markets are the non-linear, highly dynamic, and stochastic, machine learning (ML) techniques has interested researchers and practitioners since 1990s, and many articles have discussed how to use supervised learning algorithms to improve the accuracy of stock price prediction [1]. Neuro network and supervised learning methods like Decision tree, Support Vector Machine, Random Forrest etc involve training models on historical data with known outcomes, have shown significant promise in capturing market trends and predicting future stock price movements [2].

Compared to the boosting economy growth in China, the Chinese stock market presents a unique phenomenon that it has shown a downward trend since 16th October 2007 when Shanghai Composite Index reached its peak and started its downward fluctuation. The high volatility and significant non-linearity of A-share market brought great challenges to the stability and accuracy of the forecasting models [3, 4]. Meanwhile, even there are abundant historical data of A-share market, the data quality and policy influence also pose challenges on the models' construction [5,6]. There is environment for testing the application of machine learning due to its distinctive characteristics.

Among the listed companies in A-share market, Kweichow Moutai (abbreviated as Moutai), a premium liquor stock is one of the highest-priced and most stable stock. Moutai has sustained over 1000 yuan since it first broke 1000 yuan in 2019, therefore it has been regarded as one of the most representative stocks of long-term return in A-share market. Whereas CATL, a major player in the battery manufacturing and electric vehicle supply chain, symbolizes the dynamic and high-growth potential of China's burgeoning new energy sector. These two companies which represent the consumption sector and new energy sector, serve as indicators of broader market sentiment in China, while reflecting the underlying performance of their respective industries.

In this study, we aim to explore the application of supervised learning model in predicting the stock prices of Moutai and CATL, including support vector machines (SVM) by incorporating different features. Furthermore, this study will evaluate the predictive performance of these models in terms of accuracy, precision, and applicability to real-world trading strategies.

2. Literature Review

The application of supervised learning techniques has become a prominent research area to capture complex patterns in stock market prediction. The literature on machine learning

(ML) for stock forecasting covers a broad spectrum of algorithms and techniques, like such as support vector machines (SVM), decision trees (DS), random forests (RF), and deep neural networks (DNN). This section reviews relevant studies that have explored these techniques, focusing on their performance in predicting stock prices and trends.

2.1 Tree-Based Models for Stock Price Direction Prediction

Sadorsky (2022) achieved prediction accuracies exceeding 85% over forecast horizons of 8 to 20 days, by employing random forests, bagging, and extremely randomized trees (ExtraTrees), to predict stock price directions in the solar energy sector [7]. The study highlighting the effectiveness of tree-based models in handling complex, high-dimensional datasets, proved that the use of technical indicators and external economic factors (e.g., oil price volatility) can enhance the predictive performance of models. Moreover, this underscores the importance of feature selection and model combination in achieving high prediction accuracy in supervised.

2.2 Integrating Supervised Learning with Econometric Approaches

As one of the most used supervised learning techniques, support vector machines (SVM) has shown its great advantages in outperform classical econometric models when applied to financial datasets with complex structures [8,9]. Shobana and Umamaheswari (2021) review the intersection of econometrics and machine learning in financial forecasting. Their work emphasizes the increasing use of supervised learning techniques alongside traditional econometric models to enhance the predictive power of financial analyses [10], and combining variables such as interest rates, inflation, and market indices can improve the accuracy of long-term stock predictions [11]. This approach is particularly applicable to Moutai, a long-term growth stock in the Chinese market, where understanding macroeconomic trends and their impact on stock performance is essential. Their review points to the growing trend of integrating ML models with traditional financial analysis frameworks, thereby providing a more comprehensive approach to stock forecasting.

2.3 Hybrid Machine Learning Models

Zhong and Enke (2019) introduce a hybrid approach that combines deep neural networks (DNNs) with traditional artificial neural networks (ANNs) to predict the daily return direction of SPDR S&P 500 ETF. By leveraging feature selection methods such as principal component analysis (PCA), they improved the overall performance of their prediction models. The hybrid model outperformed standard ANN and DNN approaches, suggesting that combining multiple machine learning algorithms can yield better predictive results [12]. K-nearest neighbor (KNN) algorithms, an unsupervised learning model, was employed to reduce data sparsity and outlier effects to improve the classification of stock trends and improve the model robustness. The KNN classifier proved highly effective in this context, outperforming traditional models in terms of error reduction, with low mean square error (MSE) and mean absolute error (MAE) [13]. Although their study focused on the U.S. stock market, the hybrid approach is adaptable to the Chinese market, particularly for high-growth stocks like CATL, where both short-term fluctuations and long-term trends are critical for investment decisions.

2.4 Comparative Performance of Supervised Learning Models and Studies in Chinese Market

A comparative analysis of supervised learning models, including ANNs, SVMs, random forests, and naive Bayes, was conducted in 2019 to predict stock price movements in the Taiwanese market. The researchers found that ANNs performed best when predicting continuous stock prices, while SVM and random forests excelled in predicting directional movements [14]. A study conducted by Chen and Hao, established a model of a feature weighted support vector machine and K-nearest neighbor algorithms and achieved a better prediction capability to Shanghai Stock Exchange Composition Index and Shenzhen Stock Exchange Component Index in the short, medium, and long term respectively [15]. Their findings are relevant to the Chinese market, where stocks like Moutai, known for stable price trends, and CATL, with its dynamic price movements, could benefit from model selection tailored to the specific characteristics of each stock.

2.5 Conclusion

The reviewed literature demonstrates that

supervised learning techniques have proven its high effectiveness and outperform the traditional regression methods in predicting stock prices and trends, particularly when combined with feature selection and data processing methods. For stocks like Moutai and CATL in the Chinese market, with the integration of external economic factors, models such as SVMs, random forests, and hybrid neural networks offer robust solutions to capturing both short-term price fluctuations and long-term market trends.

3. Methodology

3.1 Data Collection

Two stocks, Kweichow Moutai (Maotai) and Contemporary Amperex Technology (CATL) have been selected as the target stock in this study.The dataset used in this study comes from the following authoritative organizations, including the Shanghai Stock Exchange (SSE), the Shenzhen Stock Exchange (SZSE), the National Bureau of Statistics of China (NBSC), the People's Bank of China (PBOC), and the Ministry of Industry and Information Technology of China (MIIT), with a time span from 1st January 2019 to 31st August 2024, with 1374 records.

The data contains information about the daily performance of two stocks, Shanghai Composite Index, and Shenzhen Component Index, indexes of macro-economy including GDP growth rate, exchange rate of RMB vs USD, CPI, and 10-years Government Bond Yield, the impact of unexpected black-swan, like COVID-19, and seasonal impact on the two sectors as well.

3.2 Data Processing

After filling missing values in the dataset by using forward and backward filling methods, all non-binary variables were standardized using Z-score normalization to ensure comparability and to improve the performance of machine learning models. The dataset was split into training (80%) and testing (20%) sets based on a time-series approach rather than random samplying. The earliest 80% of the dataset was assigned to the training set, and the rest 20% of the dataset was assigned to test set, so that the distribution of stock prices in the training and testing sets was consistent, and there is no overlapping data to prevent inadvertent information leakage. Table 1 presents the response variables, and the features variables.

Table 1. Response Variables and Feature Variables

Response Variables	Close prices of Maotai and CATL
Features	5-day, 10-day, and 20-day simple moving averages (SMA) of 2 stocks daily volums of 2 stocks daily close prices and volumes of Shanghai Composite Index and Shenzhen Component Index GDP growth rate CPI Exchange rate RMB vs USD 10-year Government Bond Yield Impact of COVID-19 Seasonal impact of Baijiu consumption which is related to Maotai Seasonal impact of EV consumption which is related to CATL Wheat future price and volume which are related to Maotai EV penetration in China market which is related to CATL
Source	the Shanghai Stock Exchange (SSE), the Shenzhen Stock Exchange (SZSE), the National Bureau of Statistics of China (NBSC), the People's Bank of China (PBOC), the Ministry of Industry and Information Technology of China (MIIT)

3.3 Machine Learning Method

Support Vector Machines (SVM) was adopted in this study. SVM is a supervised learning model which is especially useful for datasets with a large number of features compared to the number of samples. SVM is a classification model that finds an optimal hyperplane to separate samples of different classes.

$$w \cdot x + b = 0 \quad (1)$$

$$\min_{w,b} \frac{1}{2}||w||2 + C\xi i \quad (2)$$

w: the normal vector

b: he bias term

ξi: slack variables

C: regularization parameter

We constructed Support Vector Machine (SVM) models for both Maotai and CATL stocks under three different scenarios for each stock. The

scenarios were defined as Table 2.

Table 2. Features for Models

Models	Features
Maotai Model 1	Trading volume and 5-day, 10-day, and 20-day moving averages of Maotai; close prices and trading volumes of Shanghai Composite Index and Shenzhen Component Index
Maotai Model 2	Features in Maotai Model1 + macroeconomic indicators (CPI, GDP growth rate, exchange rate, and 10-year Government Bond Yield) Wheat future price and volume Impact of COVID19
Maotai Model 3	Features in Maotai Model2 + Seasonal impact of Baijiu consumption
CATL Model 1	trading volume and 5-day, 10-day, and 20-day moving averages of CATL; close prices and trading volumes of Shanghai Composite Index and Shenzhen Component Index
CATL Model2	Features in CATL Model1 + macroeconomic indicators (CPI, GDP growth rate, exchange rate, and 10-year Government Bond Yield) EV penetration in China market Impact of COVID19
CATL Model3	Features in CATL Model2 + Seasonal impact of EV consumption

Each model was evaluated using the following metrics on the test set: MSE, RMSE, MAE, MAPE, directional accuracy. Since the output of the prediction is the price of the stocks, to evaluate the accuracy, direction accuracy was adopted to measures the proportion of days where the predicted direction of stock price change (increase or decrease) matches the actual direction.

3.4 Results

The results from the six models, three for Kweichow Moutai (Maotai) and three for Contemporary Amperex Technology (CATL or Ningde), are summarized in Table 3 as follows.

It could be told from Table 3 that for Maotai, Model 1 showed the highest direction accuracy at 74.91%, indicating that the basic stock features were most effective in predicting the correct price movement direction. The inclusion of macroeconomic and seasonal factors in Models 2 and 3 did not improve direction accuracy significantly. Moreover Model 1 had the lowest MSE and RMSE, suggesting that the simplest model configuration with only stock-specific features resulted in the most accurate predictions in terms of minimizing errors. However, Model 3 had the lowest MAPE at 14.28%, indicating that the inclusion of seasonal factors significantly reduced the relative prediction error, making it more robust to variations in stock prices.

Table 3. Models' Evaluation.

Model	Direction_Accuracy	MSE	RMSE	MAE	MAPE
maotai_model1	74.91%	0.0052	0.0722	0.0558	21.4610
maotai_model2	71.22%	0.0055	0.0745	0.0576	30.7918
maotai_model3	72.32%	0.0053	0.0730	0.0560	14.2850
CATL_model1	69.74%	0.0047	0.0684	0.0526	36.2693
CATL_model2	74.91%	0.0044	0.0660	0.0500	39.7396
CATL_model3	69.37%	0.0048	0.0690	0.0521	39.7775

For CATL, model 2 achieved the highest direction accuracy at 74.91% and lowest MSE and RMSE, MAE and MAPE. The result indicated that the inclusion of macroeconomic indicators, such as EV penetration rate, was crucial in accurately predicting CATL's price.

The prediction error might be affected by the macroeconomic data votality and the limitation in the SVM model itself. Even though the indexes of GDP, inflation rate, and exchange rate have reflected the dynamic macroeconomi, there is a lag in their impact on stock prices, which the models may not capture fully in real time. With respect to the limitation in the SVM model. SVM has difficulties capturing the complex temporal dependencies inherent in stock price movements, irregular events, such as the COVID-19 pandemic. Additionally, sector-specific indicators, like EV penetration for CATL, are only reported periodically, which may lead to stale data and misalignment with actual market conditions.

4. Discussion

The analysis revealed that while stock-specific features were effective in predicting price movements, external macroeconomic factors and seasonal indicators have different impact on

different stocks. The mixed performance of seasonal indicators suggests that their influence may vary depending on the stock and its sensitivity to market trends and external events. It is important to distinguish the stocks which are more impacted by the macroeconomic environment and contextual factors such as consumer behavior trends and global events, whereas some are not at all.

Stock price movements are influenced by multiple factors overtime, following complex and non-linear pattern, therefore, further research could focus on exploring more complex machine learning models, such as Long Short-Term Memory (LSTM) networks and Random Forests. LSTM can adapt to these patterns and potentially provide more accurate predictions by capturing the dependencies with these sequences. However traning LSTM model requires large dataset since it is computationally intensive. Radom Forests could capture the non-linear relationships between features, and rank the features by importance, which is valuable to understand which macroeconomic index or seasonal factors are the most influential factors to the stock performance. Compared to LSTM, RF is less computationally demanding and relatively easier to train and tune, with reliable predictions on long-term trends. Future research could focus on evaluating the predictive performance of these models in different forecasting horizons, to enhance both accuracy and interpretability.

References

[1] Strader, T.J., Rozycki, J.J., Root, T.H. and Huang, Y.H.J., 2020. Machine learning stock market prediction studies: review and research directions.Journal of International Technology and Information Management,28(4), pp.63-83.

[2] Sonkavde, G., Dharrao, D.S., Bongale, A.M., Deokate, S.T., Doreswamy, D. and Bhat, S.K., 2023. Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. International Journal of Financial Studies,11(3), p.94.

[3] Chen, Z., Härdle, W.K. and Jeong, S., 2018. Forecasting volatility with sentiment indicators. Journal of Forecasting, 37(8), pp.1030-1043.

[4] Li, Q., Wang, J., Wang, S. and Zhang, C., 2020. Forecasting stock prices using wavelet transform and long short-term memory neural network: An integration of time–frequency analysis and deep learning. Applied Soft Computing, 94, p.106435.

[5] Xu, Y., Hu, H., Jiang, Y. and Wang, Y., 2017. A stock selection model using sentiment analysis and machine learning techniques. Proceedings of the 2017 International Conference on Artificial Intelligence: Technologies and Applications. pp.7-12.

[6] Chen, C., Jiang, G.J. and Tong, W.H., 2021. Political uncertainty and stock market volatility: Evidence from the Chinese stock market. Journal of Financial Stability, 53, p.100806.

[7] Sadorsky, P., 2022. Forecasting solar stock prices using tree-based machine learning classification: How important are silver prices?.The North American Journal of Economics and Finance, 61, p.101705.

[8] Choudhry, R. and Garg, K., 2008. A hybrid machine learning system for stock market forecasting.International Journal of Computer and Information Engineering,2(3), pp.689-692.

[9] Shen, S., Jiang, H. and Zhang, T., 2012. Stock market forecasting using machine learning algorithms.Department of Electrical Engineering, Stanford University, Stanford, CA, pp.1-5.

[10] Shobana, G. and Umamaheswari, K., 2021, January. Forecasting by machine learning techniques and econometrics: A review. In 2021 6th international conference on inventive computation technologies (ICICT) (pp. 1010-1016). IEEE.

[11] Kyriakou, I., Mousavi, P., Nielsen, J.P. and Scholz, M., 2021. Forecasting benchmarks of long-term stock returns via machine learning.Annals of Operations Research,297(1), pp.221-240.

[12] Zhong, X. and Enke, D., 2019. Predicting the daily return direction of the stock market using hybrid machine learning algorithms.Financial innovation, 5(1), pp.1-20.

[13] Rapach, D.E. and Zhou, G., 2020. Time-series and cross-sectional stock return forecasting: New machine learning methods.Machine learning for asset management: New developments and financial applications, pp.1-33.

[14] Huang, S. and Liu, S., 2019. Machine

learning on stock price movement forecast: the sample of the Taiwan stock exchange. International Journal of Economics and Financial Issues,9(2), p.189.

[15] Chen, Y. and Hao, Y., 2017. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. Expert Systems with Applications,80, pp.340-355.

